Current Topics in Virology

Short Communication

The development and expansion of the research fronts in HIV/AIDS research: fast algorithm for detecting community structure in networks

Cebo G. Z. Daniel*

Global Center for research and Development, Berlin, Germany.

ABSTRACT

In this study the emergence, structure and dynamics of the paradigmatic research fronts that established the principles of the biomedical knowledge on HIV/AIDS are studied and described. A search of papers with the identifiers "HIV/AIDS", "Human Immunodeficiency Virus", "HIV-1" and "Acquired Immunodeficiency Syndrome" in the Web of Science (Thomson Reuters), was carried out. A citation network of those studies was formed. Then, a sub-network of the papers with the highest amount of inter-quotations (with a minimal in-degree of 28) was chosen to represent a combination of network clustering and text mining to describe the paradigmatic research fronts and evaluate their dynamics. Thirteen research fronts were identified in this sub-network. The biggest and oldest front is referred to the clinical knowledge on the condition in the patient. Nine of the fronts are related to the study of specific molecular structures and mechanisms and two of these fronts are related to the development of drugs. The rest of the fronts are related to the research of the disorder at the cellular level. Interestingly, the emergence of these fronts appeared in subsequent "waves" over the time which suggests a transition in the paradigmatic focus. The emergence and evolution of the biomedical fronts in HIV/AIDS research is illustrated not just by the separation of the problem in elements and communications leading to increasingly specialized societies, but again by changes in the technological context of this health problem and the dramatic changes in the epidemiological reality of HIV/AIDS that occurred between 2013 and 2015.

KEYWORDS: network analysis, HIV, drug research and development, evolutionary immunology, network model.

*E-mail id: daniel.cebo@t-online.de

Introduction

The Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS) is an international health problem: over 70 million people have been infected with HIV, 35 million have died and 36.7 million people presently live with the disorder [1]. HIV/AIDS is one of the most investigated infection diseases with more than 260,000 papers listed in GO PubMed [2] and more than 42,000 papers in the Web of Science [3] covering over thirty year of scientific research. HIV/AIDS is studied by a plurality of biomedical methods like epidemiology [4], virology [5], immunology [6] or drug development [7] and non-biomedical disciplines like social sciences [8] and humanities [9]. All the biomedical disciplines working on HIV/AIDS heavily rely on a solid scientific consensus, which explains the clinical expression of HIV/AIDS in terms of the virus interactions with the immune system cells; the behavior and demography of the immune system cells, and, most notably, the virus cooperation with the biomolecular machinery of the host cells [10]. Two features are at the core of the scientific consensus on HIV/AIDS: the natural history of the HIV infection (the number of CD4+ cells and

HIV RNA copies plotted over the time) [11] and the virus replication cycle (from the virus access to the virus assembly, growing and maturation) [12].

Paradigms are the keystone of research communities [13, 14], for they provide a foundation for members of the community; they further illustrate the questions, the standards, the rules and the predicted results that guide research efforts. Paradigms of HIV/AIDS research are usually presented in a timeline format [15]. However, while such a historical context is informative, they present two disadvantages: the first is that the selection of the most significant discoveries is arbitrary, i.e., not supported by scientometrics evidence, while the second disadvantage is that the paradigms are not presented as the key elements of the organizing process of the research communities.

The study of the emerging research fronts offers the opportunity of analyzing the relation between the paradigms and the regulatory process of the scientific communities [16, 17]. Research fronts can be recognized as modules or clusters in a citation network of papers, i.e., sparse sub-networks of papers that exhibit dense connections [18].

It must be pointed out that research fronts are the impression of the scientific societies. That is, citation patterns of scientists exhibit homophily [19], which is produced by the scientists' trend to refer to those studies that concentrate in related topics with a similar approach - and surprisingly generally they cite those papers that support the papers' argumentation [20]. Citations tend to point toward those data that the research (sub)communities recognize as the most significant ones. i.e., the paradigms [21-23]. Therefore, paradigms involve the most important point in the citation networks; they are the seeds that form the development of the research fronts. To explain the emergence of the biomedical consensus on HIV/AIDS involves a study of the structure and dynamics of the research fronts.

Previous investigations using the research fronts analysis procedure were mainly directed in topics from engineering [24, 25], biotechnologies and scientometrics [26]. There are some studies that focused on the structure of the biomedical knowledge on specific diseases [27, 28]. Recent research has been especially concentrated in the core region of the literature networks. Others have reported the evolution of research fronts in anthrax research, cancer research and cardiovascular medicine.

Objective

Through a combination of text mining and network analysis, this study sought to understand the emergence and evolution of the research fronts (the footprints of the research communities) that produced the paradigmatic explanation of HIV/AIDS.

Methodology

A search of papers on HIV/AIDS was performed in the Web of Science during March 2017. The search criteria were the following: TITLE: ("HIV AIDS") OR TITLE: ("Human immunodeficiency virus") OR TITLE: ("acquired immune deficiency syndrome") OR TITLE: (hiv-1). Refined by: DOCUMENT TYPES: (ARTICLE). Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI. 60,464 papers were found.

A network model was built with the papers found in the Web of Science by using the software HistCite [29]. Then, the network model was analyzed and visualized with Cytoscape [30]. The indegree distribution of the network was evaluated to determine if it fitted to a power law function ($y = ax^b$).

A core sub-network of papers with an indegree ≥ 28 was then closely examined. Normally, the indegree distribution in citation networks follows a power law function such that only a few papers are very well cited, while most papers are not [29]. This applies to the case of HIV/AIDS research as has been reported. Papers with an indegree ≥ 28 were selected because they are a small and workable quantity of papers that account for nearly half of the communication process through the citations network as it is reported in the results section (The selected papers received 42,8911 out of 679,497 citations from the HIV/AIDS literature). Top cited papers appear to be related to the paradigmatic milestones of a particular research topic.

A cluster analysis based on the Newman modularity [30] was performed on the core sub-network

using Clust&see, a Cystoscape plug-in [31]. This analysis divided the sub-network of citation into several research fronts (clusters or modules of papers). These clusters are defined by Newman as "groups of vertices within which connections are dense but between which they are sparse" [32].

The sub-network was displayed by using the "yFiles organic" algorithm, which is based on the force-directed layout [33]. This algorithm considers the nodes as charged particles that exhibit repulsive forces and the vertices as springs. In this layout, the papers that cite the same papers tend to stick together making easier the visualization of the research fronts.

The number of papers of each research front was plotted over the years in order to track the dynamics of the research fronts.

The content of the identified research fronts - the abstract of their papers - was analyzed with KH Coder [34], a software for quantitative content analysis (Text mining). KH Coder delivered several outputs. However, this study considered that the most informative output was the list of the most distinctive words which provided key information about what was the main focus of the papers of each front. Additionally, the five papers with the highest indegree within each of the research fronts were identified in order to provide a context to the reading of the text mining results.

Because front 1 "patient" is the largest and most central front according to the results, a cluster analysis was then performed on it by using Clust&see. The sub-modules that form front 1 were identified.

Results

The network models

60,464 published articles on HIV/AIDS were identified by keyword search over the Thomson Reuters Web of Science. 57,485 of these papers form a single network of 679,497 inter-citations. The structural network analysis performed by Cytoscape showed that the distribution of the indegree in this network fitted a power law function ($y = ax^b$, $a = 51{,}954$, $b = -1.79$, correlation = 0.827, R-squared = 0.909). This means that a very small number of papers receive most citations while most papers receive few if any citations [35].

Papers with an indegree ≥ 28, that is, 5,933 documents, were selected. Together, these papers receive 63% of the inter-citations that form the whole network (42,8911 of 679,497), and would represent a relevant part of the historical core of the HIV/AIDS research as it was explained in the methodology section. These 5,933 highly cited papers formed a network of 86,963 inter-citations. The cluster network analysis identified fourteen clusters (or modules as defined by Newman). Thirteen clusters were formed by 12,303, 9,115, 7,407, 6,746, 5,680, 4,763, 4,696, 3507, 2,861, 2,768, 2,597, 2,053, and 1,662 inter-citations.

The research fronts clearly can be grouped in three different periods of time in which the fronts reach their maximum number of papers per year: 1990–1991, 1996–1999 and 2004–2017. In order to properly read the results, it is important to keep in mind the dramatic changes in the epidemiology of HIV/AIDS in the United States (USA) that happened between 1993 and 1995. In that period, the number of AIDS diagnosis and deaths reached their maximum and then declined. Simultaneously, in 1995, the number of persons living with HIV began to rise [36]. Therefore, we can consider the existence of two stages in the history of HIV/AIDS: before 1995 in which AIDS was the main concern and after 1995 when HIV infection is at the center of HIV/AIDS research. A second important consideration to understand the results is that the phase of expansion or growth in science (the "normal" science of Thomas Kuhn) follows the publication of those scientific achievements that organize the subsequent research. This would explain the peaks that generally occurred years after the publication of the papers with the highest degree. The peaks can be considered a delayed response to fundamental events and discoveries in the history of HIV/AIDS research. A third consideration is that the network model is made from the ten percent of papers with the highest indegree. Therefore, the succession of research fronts observed does not mean the end of the research on specific topics but that these topics are no longer in the core of HIV/AIDS research.

Fronts 2 "glycoprotein 120", 4 "tat-tar", 5 "reverse transcriptase inhibitor" and 10 "brain" emerged immediately after front 1 and peaked in the 1990 and 1991 years. The expansion of these fronts in

this early stage in the history of HIV/AIDS research suggests that these fronts are relevant to the description, explanation or intervention of AIDS. For example, it has been pointed out that tat (Trans-activator of transcription) protein, which is essential for virus replication, could be involved in the progression to AIDS and in the development of Kaposi's sarcoma lesions [37, 38]. Along the same line, the interaction between glycoprotein 120 and CD4 is the first event in the replication cycle and is considered fundamental to virus entry [39]. It is important to keep in mind that the depletion of lymphocytes expressing CD4 is considered the most severe hematological feature of AIDS [40]. Similarly, encephalopathy is one of the most dominant features of AIDS [40]. Finally, a reverse transcriptase inhibitor, zidovudine (AZT) was the first drug approved by the United States Food and Drug Administration (FDA) to treat AIDS [41].

Research fronts 1 "patient" and 3 "isolate" reached their maximum number of papers per year between 1996 and 1999. The peaks of these fronts follow the changes in the epidemiology of HIV/AIDS in the USA. Therefore, these fronts are possibly related to a collective response from the scientific community to the new reality of the disease. The research in front 1 is the largest, central and most clinical among the fronts. This front connects the clinical and epidemiological manifestations of HIV/AIDS with their explanation at a cellular level. Because of the size, the centrality and clinical relevance, it was decided to perform a second round of cluster analysis to identify the sub-modules that may conform front 1. The papers with the highest degree are: "Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment [42], "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection" [43], and "Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy" [43]. These papers report and explain fundamental changes in the clinical reality of HIV/AIDS produced by the implementation of anti-retroviral therapies. Front 3 "isolate" is related to to the study of HIV tropism, i.e., the differential capacity of the HIV strains to infect and replicate in different cell types. Importantly, the availability of screening tools that allowed the identification of asymptomatic individuals infected with HIV and the use of anti-retroviral therapies make the blood and tissue samples extensively available from the patients that were fundamental to the emergence of front 3.

On the other hand, most of the research fronts specialized in the study of specific molecular mechanisms and structures (fronts 6 to 9 and fronts 11 to 13) peaked either in the 1996–1999 or 2004–2007 periods. It is important to notice that all these fronts emerged at the end of the first decade of HIV/AIDS research. The difference between fronts peaking in the second and the third periods is that the former decline earlier. In order to understand the evolution of these fronts it is important to keep in mind that the scientific specialization is a continuous process of solving problems that follows the establishment of a paradigm (HIV-1 as the etiological agent of HIV/AIDS). In that sense, the decline of fronts 8, 11 and 12 may be so because the scientific problem is either essentially solved or the changes in the HIV/AIDS epidemiology made the topics related to these fonts less relevant.

Discussion

According to the Cytoscape analysis, the network of HIV/AIDS papers displays a power law distribution of their citations, which has important methodological implications. The first implication is that a research front (a citation network module) could be formed by other research fronts, which in turn can be portioned into sub-modules [44]. The second implication is that the nodes (papers) with the highest indegree tend to be more "cosmopolitan" i.e., they have the lowest clustering coefficient values. That is, they could belong simultaneously to several fronts or any of them. Therefore, there are not clearly defined frontiers dividing the research fronts. However, the standard in the scientometrics study of research fronts seems to be to use clustering methods that define frontiers between the modules, probably because this makes the community structure in a literature network much more understandable. Moreover, analysis reveals the front that is most relevant to each paper. Finally, the most important implication is that in a hierarchical literature network most of the papers with the highest indegree are related with the

paradigms that organize a research field or topic. Top cited papers have been extensively used to identify the scientific achievements that establish the standards of research practice of a community [45, 46]. There are no set guidelines on the proportion of top cited papers that should be selected. However, there is a tradeoff between selecting the most informative papers and maintaining diversity of the information. In this work, a minimal indegree of 30 was used to select the top cited papers, and in turn obtaining a considerable percentage of the citations. The selected papers consist of only ten percent of the network, but they effectively account for two thirds of the citations. The selected papers are a reasonable representation of the paradigmatic core of HIV/AIDS research.

Once the paradigms are established, researchers focus on the details, the smaller range of problems and solutions that the current paradigm provides Results suggest that the emergence of several of the specialized research fronts was caused by the partition of the general problem in interacting elements. That is, HIV/AIDS research could be understood to some extent as an instance of part-whole science in which paradigms determine the abstraction of the parts that are considered the most relevant to explain the whole phenomenon [47, 48].

The general structure and evolution of the research fronts in HIV/AIDS research shares similarities to that of anthrax and Ebola. The evolution of anthrax investigation began with a preliminary on the immunology of the disease. From this, four research fronts emerged: “anthrax gene sequencing”, “vaccine research”, “secondary research on PA (protective antigen) and LF (lethal factor)”, and “making and purifying toxin”. Subsequently, the research front on PA and LF split in three fronts: “specific PA research”, “PA-mediated delivery of other substances” and specific “LF research”. Similarly, the evolution of the fronts in Ebola research are marked by a front related to the report of the epidemiology and the clinical manifestation of the disease. A second front provides an explanation of the disease at tissue-cellular level. Then, research on Ebola split into four research fronts, each one specialized in one different virus protein. There is also a front aimed to the development of vaccines and other immunotherapies. Similarly, the emergence of the fronts in HIV/AIDS research started with a general research front that provided the pathology of the disease and subsequently split into specialized fronts focused on the study of specific molecular mechanism of the virus replication cycle. In all three cases, the specialization of the research led to the emergence of research fronts focused in the study of the parts that are thought to be key in explaining the diseases. A report on the emergence of the research fronts in cancer and cardiovascular diseases showed that the specialization process in these types of diseases is complex. Jones *et al*. reported fronts specialized on microarrays, targeted therapies, clinical trials, epidemiology and molecular etiology in cancer research, while in cardiovascular diseases the fronts are organized around drug-eluting stents, anti-platelet agents, pacemakers, hypertension and atrial fibrillation. The difference between these two groups of diseases is that HIV/AIDS, anthrax and Ebola are infectious diseases with a clearly identified etiological agent while cancer and cardiovascular diseases are both complex multifactorial diseases [49].

Conclusion

This is the first time that the complex organization (and the evolution) of HIV/AIDS research is reported. This research provides fundamental knowledge concerning the emergence of the paradigmatic explanation for HIV/AIDS and therefore contributes to the understanding of the nature of biomedical knowledge. In addition, this work suggests that the development of the paradigmatic knowledge on HIV/AIDS in terms of the emergence and evolution of the research fronts followed two different routes. First, the emergence of the specialized fronts (molecular mechanism and structures and cellular process) was caused by the division of the general problem in their key process, element and interactions, which is related to the concept of part-whole science. Second, the dynamics of the fronts, particularly the evolution of front 1 “patient” and 2 “isolate”, appears to represent an adaptive and collective response from the scientific community to changes in the epidemiological (the decline in the morbidity and mortality of AIDS in the USA) and technological (the availability of treatments and screening tools) context of this health problem.

CONFLICT OF INTEREST STATEMENT

This paper does not contain any conflict of interests.

REFERENCES

1. World Health Organization. Global health observatory (GHO), 2015, URL: http://www.who.int/gho/tb/en
2. Doms, A. and Schroeder, M. 2005, Nucleic Acids Research, 33(Suppl. 2), W783-6.
3. Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. and Pappas, G. 2008, FASEB J., 22(2), 338-42.
4. Wilson, D. 2006, HIV epidemiology: A review of recent trends and lessons. The World Bank: Washington DC.
5. Sierra, S., Kupfer, B. and Kaiser, R. 2005, Journal of Clinical Virology, 34(4), 233-44.
6. Medzhitov, R. and Littman, D. 2008, Nature, 455(7213), 591.
7. Flexner, C. 2007, Nature Reviews Drug Discovery, 6(12), 959-66.
8. Gillies, P. A. 1996, AIDS in the World II: Global dimensions, social roots, and responses, 131-58.
9. Cebo, D. 2017, Current Trends in Immunology, 18, 85-89.
10. Fauci, A. S. 2003, Nature Medicine, 9(7), 839-43.
11. Kallings, L. O. 2008, Journal of Internal Medicine, 263(3), 218-43.
12. Barré-Sinoussi, F., Ross, A. L. and Delfraissy, J. F. 2013, Nature Reviews Microbiology, 11(12), 877-83.
13. Cebo, D. 2017, Current Topics in Virology, 14, 59-67.
14. Fajardo, D. and Castano, V. M. 2016, Curr. Med. Chem., 26, 3000-3012.
15. Merson, M. H. 2006, New England Journal of Medicine, 354(23), 2414-7
16. Morris, S. A., Yen, G., Wu, Z. and Asnake, B. 2003, Journal of the American Society for Information Science and Technology, 54(5), 413-22.
17. Fajardo-Ortiz, D., Ortega-Sánchez-de-Tagle, J. and Castaño, V. M. 2015, Journal of Translational Medicine, 13(1), 1.
18. Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I. and Matsushima, K. 2011, Technol Forecast Soc Change, 78(2), 274-82.
19. Ding, Y. 2011, Journal of Informetrics, 5(1), 187-203.
20. Greenberg, S. A. 2009, BMJ, 339, b2680.
21. Baltussen, A. and Kindler, C. H. 2004, Intens. Care Med., 30(5), 902-910.
22. Hennessey, K., Afshar, K. and Macneily, A. E. 2009, Can. Urol. Assoc. J., 3(4), 293-302.
23. van Noorden, R., Maher, B. and Nuzzo, R. 2014, Nature, 514(7524), 550-553.
24. Shibata, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. 2008, Technovation, 28(11), 758-75.
25. Shibata, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. 2009, Journal of the American Society for Information Science and Technology, 60(3), 571-80.
26. Anegón, F. D, Contreras, E. J. and Corrochano, M. D. 1998, Scientometrics, 42(2), 229-46.
27. Fajardo-Ortiz, D., Ochoa, H., Garcia, L. and Castano, V. 2014, Cadernos de Saúde Pública., 30(2), 415-26.
28. Jones, D. S, Cambrosio, A. and Mogoutov A. 2011, Journal of Translational Medicine, 9(1), 1.
29. Garfield, E. 2009, Informetrics, 3(3), 173-179.
30. Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A. Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P. L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T. and Bader, G. D. 2007, Nat. Protoc., 2(10), 2366-82.
31. Brzezinski, M. 2015, Scientometrics, 103(1), 213-228.
32. Newman, M. E. 2004, Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 69(6 Pt 2), 066133.
33 Spinelli, L., Gambette, P., Chapple, C. E., Robisson, B., Baudot, A., Garreta, H., Tichit, L., Guénoche, A. and Brun, C. 2013, Biosystems, 113(2), 91-95.
34. Higuchi, K. 2012, Shakai Chosa., 8, 92-6 (In Japanese).
35. Clauset, A., Shalizi, C. R. and Newman Mark, E. J. 2009, SIAM Rev., 51(4), 661-703.

36. Centers for Disease Control and Prevention (CDC). HIV surveillance-United States, 1981-2008. 2011, MMWR Morb Mortal Wkly Rep., 60(21), 689-93.
37. Taylor, J. P., Pomerantz, R., Bagasra, O., Chowdhury, M., Rappaport, J., Khalili, K. and Amini, S. 1992, EMBO J., 11(9), 3395-403.
38. Ensoli, B., Barillari, G., Salahuddin, S. Z., Gallo, R. C. and Wong-Staal, F. 1990, Nature, 345(6270), 84-6.
39. Lyerly, H. K., Matthews, T. J., Langlois, A. J., Bolognesi, D. P. and Weinhold, K. J. 1987, Proc. Natl. Acad. Sci. USA, 84(13), 4601-5.
40. Price, R. W., Brew, B., Sidtis, J., Rosenblum, M., Scheck, A. C. and Cleary P. 1988, Science, 239(4840), 586-92.
41. Greene, W. C. 2007, Eur. J. Immunol., 37(Suppl. 1), S94-102.
42. Connor, E. M., Sperling, R. S., Gelber, R., Kiselev, P., Scott, G., O'sullivan, M. J., VanDyke, R., Bey, M, Shearer, W., Jacobson, R. L. and Jimenez, E. 1994, New England Journal of Medicine, 331(18), 1173-80.
43. Ho, D. D., Neumann, A. U., Perelson, A. S. and Chen, W. 1995, Nature, 373(6510), 123.
44. Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisson, R. E., Quinn, T. C., Chadwick, K., Margolick, J., Brookmeyer, R., Gallant, J., Markowitz, M., Ho, D. D., Richman, D. D. and Siliciano, R. F. 1997, Science, 278(5341), 1295-1300.
45. van Noorden R., Maher, B. and Nuzzo, R. 2014, Nature, 514(7524), 550-553.
46. Harris, J. K. 2010, Am. J. Public Health, 100(7), 1245-1249.
47. Winther, R. G. Synthese, 178(3), 397-427.
48. Okwundu, C. and Okoromah, C. A. 2009, Cochrane Database Syst Rev., 1, CD007189.
49. Kiberstis, P. and Roberts, L. 2002, Science, 296(5568), 685.